Causal Inference with Graphs:

- If testing effect of setting $X=x$,
  1) remove all arrows into X and leave all outgoing arrows
  2) Fix $P(X=x)=1$
  3) Calculate joint distribution in this new graph

- For ~~[illegible]~~ variables with no parents,
  $$P(Y|set(X=x)) = P(Y|X=x)$$

We want to find $P(Y|set(X=x))$ (or maybe $E[Y|set(X=x)]$)

- $P(Y|set(X=x))$ is <u>identified</u> if we can write it as a function of the joint distribution

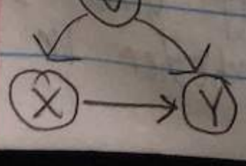
Easy Cases where $P(Y|set(X=x))$ is identified:

i) X is experimentally controlled
ii) randomize X (see if dist has no corr with other causes) $\Rightarrow P(Y|set(X=x)) = P(Y|X)$
iii) X is exogenous (has no parents in graph) $\Rightarrow P(Y|set(X=x)) = P(Y|X)$

If X is endogenous with causal ancestors:

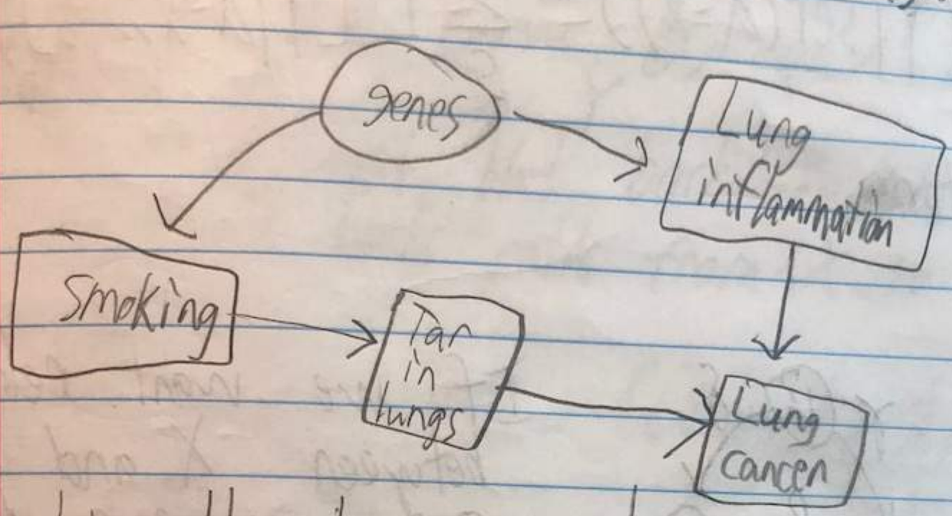
- then in general $P(Y|X=x) = P(Y|set(X=x))$

The effect of X on Y is <u>confounded</u> if X and Y share ancestors



- Merchants of Doubt (book about statistical effects and psychology)

Ex:



smoking $\perp\!\!\!\perp$ lung cancer | tar, inflammation (because all paths are blocked)

So cancer ~ smoking + tar + inflammation will have slope coeff. for smoking as 0.

smoking $\not\!\perp\!\!\!\perp$ Cancer | inflammation
(blocks path through genes to prevent confounding)

Confounding prevention:

- Don't control for descendants of X
- Control for variables that result in path from X's ancestors to Y
- Path from X to Y is a backdoor path if it has an arrow into X

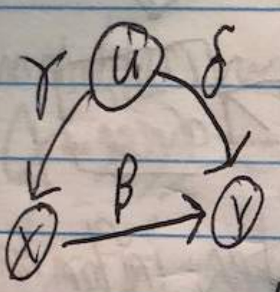
Set of variables meets back-door criterion for finding $P(Y|set(X=x))$ if: S blocks all backdoor paths and S has no descendants of X.

Then, $P(Y|Set(X=x)) = \sum_s P(Y|X=x, S=s)P(S=s)$

and $E(Y|Set(X=x)) = \sum_s E[Y|X=x, S=s]P(S=s)$

Controlling for

Ex: with confounding



If we want relation between X and Y and we run linear regression we will estimate slope as

$$\frac{Cov(X,Y)}{Var(X)} = \frac{\beta Var(X) + \gamma\delta Var(U)}{Var(X)}$$

$$= \beta + \gamma\delta\frac{Var(U)}{Var(X)}$$

so $E[Y|X=x] = \beta_0 + \left(\beta + \gamma\delta\frac{Var(U)}{Var(X)}\right)x$

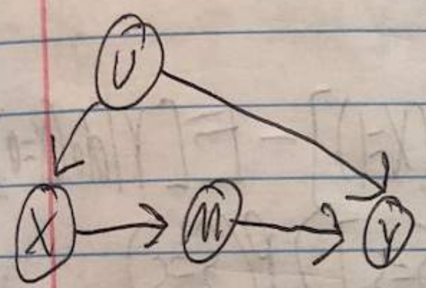
- we don't want since it is the confounding from U and not the true influence of X.

$$\sum_s P(Y|X=x, S=s)P(S=s) = E[P(Y|X=x, S)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} P(Y|X=x, S=s_i)$$

Front door approach:

- if we don't know U, and we want the effect of X on Y
  - X must block backdoor paths from M to Y
  - M blocks paths from X to Y

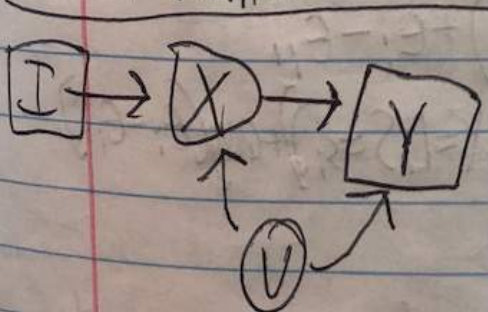


then $P(Y|do(X=x))$

$$= \sum_m P(Y|M=m, do(X=x)) P(M=m|do(X=x))$$

$$= \sum_m P(Y|do(M=m)) P(M=m|X=x)$$

Instrumental variables:



I is exogenous, an ancestor of Y, and all directed paths from I to Y go through X.

$P(Y|do(I)) = P(Y|I)$

$P(X|do(I)) = P(X|I)$

- If we have estimate of $\hat{X} = \alpha I$ (some function of I), then if we want to do a regression of Y from X, there's no confounding from U. (If we don't use X as function of I, then there is confounding from U since the path from I to Y is collider that's only open when we condition of X)

• Sometimes IV's don't work if everything is not linear

Matching:

Average treatment effect (ATE) $= E[Y|do(X=1)] - E[Y|do(X=0)]$

$$= \sum_s \left(E[Y|X=1, S=s] - E[Y|X=0, S=s]\right) P(S=s)$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}(X=1, S=s_i) - \hat{\mu}(X=0, S=s_i)\right)$$

• Suppose we can match each unit with $x_i=1$ to another unit with $x_i=0$ and the same S_i

$$Y_i - Y_{i'} = \mu(X=1, S=s_i) - \mu(X=1, S=s_i) + \epsilon_i - \epsilon_{i'}$$

$$\text{avg}(Y_i - Y_{i'}) = \text{avg}(\mu(X=1, S=s_i) - \mu(X=1, S=s_i)) + \text{avg}(\epsilon_i - \epsilon_{i'})$$

$$= ATE + \text{avg}(\text{noise})$$

Problems:

• exact matches are hard to find (could use approximate matches which is basically nearest neighbors)

• This only works if S meets the backdoor criterion

Finding Graph: (if we have p features how many possible graphs?)

for non directed graphs: num $= 2^p$ prevnum

1. Actual scientific/practical knowledge
2. Guess and test:
   - make up graph and check if data supports the independence relationships in graph
   - to test $X \perp\!\!\!\perp Y \mid S$, test if $P(X, Y \mid S) = P(X \mid S) P(Y \mid S)$
     or $P(Y \mid S) = P(Y \mid S, X)$
3. Consistent discovery:
   - automate search over DAGs and guarantee they converge on the correct answer
   - Spirtes–Glymour–Scheines (SGS) algorithm:
     - all variables are observed, data is IID, and you have a good conditional independence test
     - start with a complete and undirected graph for all variables
       1) if $V_1 \perp\!\!\!\perp V_2$, remove edge between them
       2) if $V_1 \perp\!\!\!\perp V_2 \mid V_3$, remove edge between V_1, V_2
       3) if $V_1 \perp\!\!\!\perp V_2 \mid \{V_3, V_4\}$, remove edge $V_1 - V_2$
       4) stop when we run out of variables

Condition on everything $\{V_3, \ldots, V_p\}$

- left with undirected graph
- Look for colliders: ($X - Y - Z$ is collider if $X \not\perp\!\!\!\perp Z \mid \{S, Y\}$ for all possible S) and use this to orient other edges

<u>Thm:</u> If the error rate of your conditional independence test $\to 0$ as $n \to \infty$ then estimated graph $\to$ true graph as $n \to \infty$

- By just checking for colliders in skeleton, ~~and seeing what elements are [illegible]~~ we are able to converge on the equivalence class of graphs (based on dependencies between variables)

- Admit uncertainty in DAGs and you can report estimates from different DAGs